10-21-2004

Finishing the euchromatic sequence of the human genome

Zahra Abdellah
Wellcome Sanger Institute

Alireza Ahmadi
Wellcome Sanger Institute

Shahana Ahmed
Wellcome Sanger Institute

Matthew Aimable
Wellcome Sanger Institute

Rachael Ainscough
Wellcome Sanger Institute

See next page for additional authors

Recommended Citation

Abdellah, Z., Ahmadi, A., Ahmed, S., Aimable, M., Ainscough, R., Almeida, J., Almond, C., Ambler, A., Ambrose, K., Ambrose, K., Andrew, R., Andrews, D., Andrews, N., Andrews, D., Apweiler, E., Arbery, H., Archer, B., Ash, G., Ashcroft, K., Ashurst, J., Ashwell, R., Atkin, D., Atkinson, A., Atkinson, B., Attwood, J., Aubin, K., Auger, K., Avis, T., Babbage, A., Babbage, S., Bacon, J., Bagguley, C., Bailey, J., & Baker, A. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431* (7011), 931-945. https://doi.org/10.1038/nature03001

Authors

Zahra Abdellah, Alireza Ahmadi, Shahana Ahmed, Matthew Aimable, Rachael Ainscough, Jeff Almeida, Claire Almond, Andrew Ambler, Karen Ambrose, Kerrie Ambrose, Robert Andrew, Daniel Andrews, Neil Andrews, Dan Andrews, Eva Apweiler, Hazel Arbery, Beth Archer, Gareth Ash, Kevin Ashcroft, Jennifer Ashurst, Robert Ashwell, Deborah Atkin, Andrea Atkinson, Barry Atkinson, John Attwood, Keith Aubin, Katherine Auger, Terry Avis, Anne Babbage, Sarah Babbage, Joanne Bacon, Claire Bagguley, Jonathan Bailey, and Andrew Baker

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a highly accurate sequence of the vast majority of the euchromatic portion of the human genome. The initial work followed a two-pronged approach: (1) the mapping of the human and mouse genomes[1–9] to allow the study of inherited disease and provide a crucial scaffold for genome assembly; and (2) the sequencing of organisms with smaller, simpler genomes[10–14] to serve as a testbed for method development and assist in interpreting the human genome. With success along both paths, the sequencing of the human genome itself eventually became feasible. The International Human Genome Sequencing Consortium (IHGSC), an open collaboration involving twenty centres in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC[15] and Celera Genomics[16] each reported draft sequences providing a first overall view of the human genome. These sequences allowed systematic study of the human genome itself, including identification of genes, combinatorial architecture of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphism and relationship between genetic recombination and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly accelerated biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, omitted ~10% of the euchromatic genome; it was interrupted by ~150,000 gaps; and the order and orientation of many segments within local regions had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of, at most, one event per 10^4 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with the only gaps being those refractory to all available techniques[17] (see http://www.genome.gov/10000923). The goal was challenging because the human genome is replete with such features as dispersed repeats and large segmental duplications, which greatly complicate the determination of genome structure and sequence. In fact, near-complete sequences have been obtained so far only for three multicellular organisms: the nematode[13], mustard weed[18] and the fruitfly[19]. These genomes are all roughly 30-fold smaller than the human genome and have much simpler structure.

We describe here the results of a multiyear effort by the IHGSC towards the goal of a complete human sequence. The number of gaps has been reduced 400-fold to only 341, most of which are associated with segmental duplications and will require new methods for resolution. The assembled near-complete genome sequence has an error rate of only ~1 event per 100,000 bases; it contains 2.85 billion nucleotides and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it; examines the accuracy and completeness of the sequence; and illustrates biological analyses made possible by the sequence. We do not attempt here a comprehensive analysis of the contents of the human genome. An initial analysis was previously reported[15] and a series of papers is being written describing the individual chromosomes[17,20–30], including annotation of genes and other features.

Current genome sequence

Finishing process

The process of converting the initial draft sequence into a near-complete sequence is referred to as 'finishing'. It is a complex iterative process that proceeds simultaneously at multiple scales, ranging from single nucleotides to the integrity of whole chromosomes. The fundamental challenge is that genomic regions that are not well represented or readily resolved through random shotgun sequencing tend to be highly enriched in problematic sequences. Resolving such regions required the development of special approaches, which evolved substantially over time and varied among centres.

Broadly, the finishing process involved two distinct components: (1) producing finished maps, consisting of continuous and accurate paths of overlapping large-insert clones spanning the euchromatic region of each chromosome arm; and (2) producing finished clones, consisting of continuous and accurate nucleotide sequence across each large-insert clone. In practice, these two components were tightly intertwined in that progress in each often depended on results from the other. The components are described in Boxes 1 and 2. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Note 1) and at http://www.genome.gov/10000923.

In total, we generated a shotgun sequence from 59,208 large-insert clones (total length ~5.84 gigabases (Gb)) and finished the sequence from 45,742 of these clones (total length ~3.67 Gb). The clones consisted primarily of bacterial artificial chromosomes

(BACs), but also included some P1-artificial chromosomes (PACs), yeast artificial chromosomes (YACs), fosmids and cosmids; they carried DNA from multiple anonymous sources[15]. We then chose a 'clone tiling path' of 26,720 overlapping clones across the genome, selected a 'sequence tiling path' of directly adjacent, non-overlapping segments from consecutive clones and concatenated these segments to create a near-complete genome sequence. Contributions of the IHGSC centres to this finishing phase are shown in Table 1.

Genome sequence

The human sequence reported here consists of 2,851,330,913 nucleotides, lying almost entirely within the euchromatic portion of the genome (Table 2). It is interrupted by only 341 gaps, of which 33 gaps (totalling ~198 megabases (Mb)) reflect heterochromatin, which was not targeted by the HGP, and 308 gaps (totalling ~28 Mb) are euchromatic. The euchromatic genome is thus ~2.88 Gb and the overall human genome is ~3.08 Gb. The long-range continuity of the current genome sequence is high by various measures (Table 3). The N50 length is 38.5 Mb and the *N*-average length is 40.9 Mb; these values are ~1,000-fold larger than the size of a typical human gene. (The first statistic is the length x such that at least 50% of nucleotides lie in a continuous segment of length $\geq x$, whereas the second is the average length of the contiguous segment containing a randomly chosen nucleotide.) Focusing on individual

Box 1
Finishing the physical map

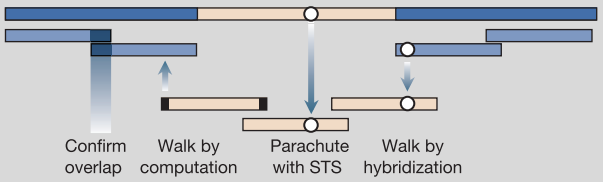
The hierarchical strategy used two kinds of genome maps as a foundation for producing finished sequence: sequence-tagged site (STS) maps[6] and clone maps[7]. The first provided global landmarks by positioning tens of thousands of STSs through genetic mapping, radiation hybrid mapping and STS-content mapping. The second provided overlapping clones for sequencing and was obtained by comparing restriction-digest fingerprints of hundreds of thousands of BAC clones to create local contigs. The two maps were integrated by anchoring the contigs to the global landmarks.

The initial random phase generated a physical map covering 96–98% of the euchromatic genome in ~1,000 anchored contigs separated by gaps with average size of ~100 kb[7]. Overlapping clones were chosen from this map to produce the draft sequence. The subsequent finishing phase involved verifying clone overlaps and closing gaps in the initial map. The finishing process for each chromosome was overseen by a designated centre and managed by a dedicated coordinator, who integrated and reviewed information from contributing centres.

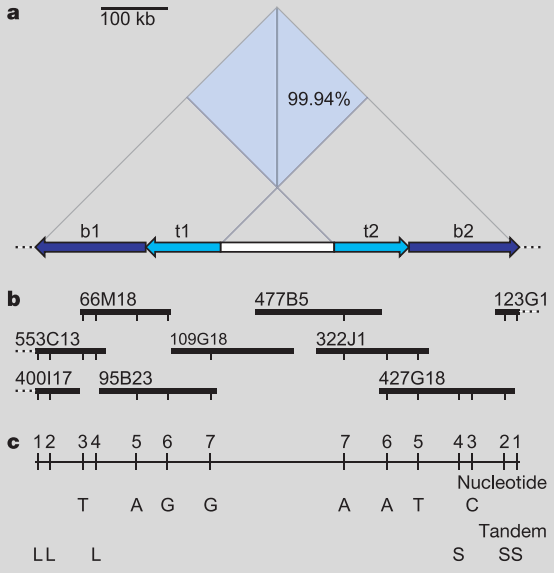
Clone overlaps were verified by analysing finished sequence to confirm that the terminal sequence overlapped for ≥2 kb with ≥99.6% identity. (Perfect identity was not expected in all cases, because the clones might derive from different haplotypes and thus differ at polymorphic sites.) A small number of overlaps ($n = 308$) falling below this threshold were accepted, because the sequencing centre presented additional evidence that the overlap was correct. For example, smaller overlaps might be confirmed by a partially sequenced clone spanning the overlap, or a region of high variation might be shown to be due to allelic variation. All these exceptions are recorded in electronic tags shown on the UCSC genome browser. Remaining overlaps were rejected and counted as new gaps in the map.

Closing gaps required identifying a tiling path of clones spanning the region. Methods used are summarized in Box 1 Fig. 1. The primary approach was iterative 'walking' from the ends of contigs: a stretch of terminal sequence from a terminal clone was used to identify a new clone extending the contig. The procedure was repeated until it either yielded a clone connecting to the neighbouring contig or reached a dead end. Walking used both experimental (hybridization to arrayed BAC libraries) and computational (comparison with a database of BAC end-sequences) methods to identify candidate clones. Also, it was sometimes possible to 'parachute' into gaps by identifying sequences likely to lie within the gap (such as STSs, mRNAs and mouse sequences in regions of conserved synteny) and using them as hybridization probes for BACs.

Some gaps could not be closed, either because no new clone could be found despite screening of diverse and deep libraries (>30-fold physical coverage) or because a complex collection of clones was found whose relationship proved impossible to discern owing to the presence of extensive highly repetitive sequences or near-exact segmental duplication. Telomeric regions of chromosomes were obtained by using a specialized cloning system ('half-YACs vectors') in which successful propagation required that the human insert DNA provided telomeric function. By far, the most difficult regions of the genome were those containing near-exact segmental duplications. A particularly challenging example is shown in Box 1 Fig. 2.



Box 1 Figure 1 Simplified flowchart for finishing the physical map.



Box 1 Figure 2 Illustration of a challenging region on chromosome Y. **a**, The sequence organization of the palindromic repeat P3 is represented by the horizontal bar at the base of the triangle. Arrows indicate orientation of sequences in the 283-kb arms of the palindrome (b1, t1 on left and t2, b2 on right). A non-repeated 170-kb 'spacer' (white) separates the arms. Above the horizontal bar, each dot represents a perfect match of 500 bp. The near identity between the arms (99.94%) appears as a vertical line of dots, highlighted by the blue diamond. **b**, Tiling path of BACs from RPCI-11 library. **c**, Sequence differences, numbered 1–7, between arms of P3. Differences 3, 5, 6 and 7 are single nucleotide differences, whereas 1, 2, and 4 are differences in the lengths of simple tandem repeats (microsatellites); L, long variant; S, short variant. These differences allowed assignment of BACs to the correct arm of P3.

Table 1 **Bases sequenced in Build 35**

Centre	Finished sequence totals (kb)	Finished sequence in Build 35 (kb)
SC	919,388	849,650
WUGSC	645,062	583,032
WIBR	562,096	373,760
JGI/SHGC	485,085	313,988
BCM	320,735	280,963
RIKEN	155,769	112,047
UWGC	145,745	105,573
GS	99,970	78,467
GTC	45,710	32,972
UWMSC	39,227	28,367
Keio	44,905	20,780
IMB	73,677	20,053
Beijing	38,079	17,114
MPIMG	9,838	5,673
GBF	8,325	5,547
UOKNOR	18,657	5,311
TIGR	10,390	3,054
CGM	2,768	1,766
SDSTDC	7,792	1,403
UTSW	8,555	196
Other	30,745	11,621
Total	3,672,516	2,851,336

Columns indicate total finished sequence deposited in public databases (including overlaps and alternative alleles) and finished sequence incorporated into Build 35 (consisting of those clones chosen for inclusion in the tiling path by the individual chromosome coordinators). The total includes finished sequence completed at the time of the draft sequence[15,17,20] and subsequently. Some clones were sequenced to draft coverage by one centre, then finished by another. Abbreviations: SI, Wellcome Trust Sanger Institute (Hinxton, UK); WUGSC, Washington University Genome Sequencing Center (St Louis, USA); JGI, US Department of Energy Joint Genome Institute (Walnut Creek, USA); WIBR, Whitehead Institute for Biomedical Research (Cambridge, USA); BCM, Baylor College of Medicine Human Genome Sequencing Center (Houston, USA); RIKEN, RIKEN Genomic Sciences Center (Yokohama, Japan); UWGC, University of Washington Genome Center (Seattle, USA); GS, Genoscope (Evry, France); Keio, Keio University School of Medicine (Tokyo, Japan); GTC, Genome Therapeutics Corporation (Waltham, USA); UWMSC, Institute for Systems Biology Multimegabase Sequencing Center (Seattle, USA); IMB, Institute of Molecular Biology (Jena, Germany); Beijing, Beijing Genomics Institute, Beijing, China); UOKNOR, University of Oklahoma's Advanced Center for Genome Technology (Norman, USA); SHGC, Stanford Human Genome Center (Stanford, USA); MPIMG, Max Planck Institute for Molecular Genetics (Berlin, Germany); GBF, German Research Center for Biotechnology (Braunschweig, Germany); TIGR, The Institute for Genome Research (Rockville, USA); CGM, Center for Genetics in Medicine (Perkin Elmer/Washinton Univ.); SDSTDC, Stanford DNA Sequencing and Technology Development Center (Stanford, USA); UTSW, University of Texas, Southwestern Medical Center (Dallas, USA). 'Other' includes clones from groups that deposited <1 Mb of finished sequence that comprises Build 35. The total includes 5,387 bp, distributed across nine chromosomes, that were ambiguous (scored as 'N') and therefore not counted in the total figure used in the text.

chromosome arms, the N50 length exceeds half the length of the arm in three-quarters of cases (Table 3).

The sequence is denoted as NCBI Human Build 35 (May 2004), with the individual chromosomes having accession numbers NC000001 to NC000024 (see Supplementary Information Note 3 concerning additional sequence data). The analyses reported here were performed on Build 35 or, in a few cases, its immediate predecessor, Build 34 (which differed only slightly). The poster accompanying this paper displays the 24 human chromosomes, together with various biological annotations. These include GC content, repeat content, segmental duplications, protein-coding genes, sequence similarity and synteny conservation with mouse, sequence similarity with the pufferfish, and density of single-nucleotide polymorphisms in the human genome. Many additional annotations can be found on public genome browsers (http://genome.ucsc.edu/; http://www.ensembl.org/; http://www.ncbi.nlm.nih.gov/genome/guide/human/), which are regularly updated.

Comparison with draft sequence

The near-complete sequence is a great improvement over the earlier draft sequence. It has substantially fewer gaps (341 versus 147,821) and greater continuity (38,500 kilobases (kb) versus 81 kb for N50 contig size), reflecting an overall improvement of ~475-fold. The draft sequence contained regions in which the local order and orientation were unknown; these have now been resolved. The case of chromosome 7 is illustrated in Fig. 1. Additionally, the draft sequence contained substantial artefactual duplication, including local events caused by errors in merging some adjacent BAC-based sequences, made by the first-generation global assembly program, and global events caused by contamination of shotgun assemblies of some BACs with data from other clones. These artefacts have now been eliminated.

Accuracy and completeness

Because the human genome sequence is intended to serve as a

Table 2 **Finished sequence and gaps, HGSC Build 35**

Chr	Total finished sequence* (kb)	Euchromatic gaps†		Heterochromatic gaps‡		Estimate of total gap size§ (kb)	Unfinished clones‖	
		Number	Est. size (kb)	Number	Est. size (kb)		Number	Est. size (kb)
1	222,828	32	1,605	2	19,510	21,115	17	850
2	237,503	20	2,512	1	2,900	5,412	0	0
3	194,636	5	1,935	1	1,500	3,435	0	0
4	187,161	14	1,250	1	3,000	4,250	0	0
5	177,703	5	92	1	340	432	0	0
6	167,318	10	658	1	2,300	2,958	0	0
7	154,759	11	869	1	4,630	5,499	0	0
8	142,613	9	662	1	2,190	2,852	0	0
9	117,781	40	1,955	2	18,000	19,955	12	600
10	131,614	12	1,020	1	2,515	3,535	8	400
11	131,131	7	322	1	4,760	5,082	0	0
12	130,259	8	795	1	4,300	5,095	0	0
13	95,560	6	715	2	17,200	17,915	0	0
14	88,291	1	8	2	17,220	17,228	0	0
15	81,342	10	737	2	18,260	18,997	0	0
16	78,885	4	143	2	10,000	10,143	0	0
17	77,800	9	875	1	7,500	8,375	0	0
18	74,656	3	97	1	1,368	1,465	0	0
19	55,786	5	5,015	1	340	5,355	0	0
20	59,505	4	1,157	1	1,766	2,923	0	0
21	34,170	3	53	2	11,620	11,673	0	0
22	34,765	11	460	2	14,330	14,790	0	0
X	150,394	12	750	1	3,000	3,750	14	700
Y	24,872	9	1,480	2	31,618	33,098	7	350
Total	2,851,331	250	25,165	33	200,167	225,332	58	2,900

*The total length of tiling paths including only finished bases of clones in Build 35. Roughly 2.19 Mb of sequence on chromosome Y was derived directly from the equivalent pseudoautosomal region on chromosome X.

†Defined as gaps in euchromatic regions, including junctions with heterochromatic/centromeric sequences, for which no clone was available (see text).

‡Defined here as gaps in heterochromatic regions (see text and Supplementary Note 2 on heterochromatic sequence). Separate gaps were counted for centromeres and pericentric heterochromatin, even when the two were contiguous. Centromere sizes were taken from ref. 62 or in some cases provided directly by the sequencing centres (see Supplementary Note 2). Acrocentric sizes are based on centromere ratios from ref. 63. The sizes of large heterochromatic gaps are typically difficult to estimate accurately owing to their repeat structure and polymorphic nature[62,64]. Other regions might arguably be called heterochromatin (for example, the pericentric regions of chromosomes 19 and 3 and a ~400-kb gap on the Y chromosome[23]), but are classified as euchromatin here.

§The sum of lengths for finished sequence, estimated heterochromatic gaps, euchromatic gaps and unfinished clone gaps. The total length is only approximate because of uncertainty in gap sizes, particularly for heterochromatic gaps and centromeres.

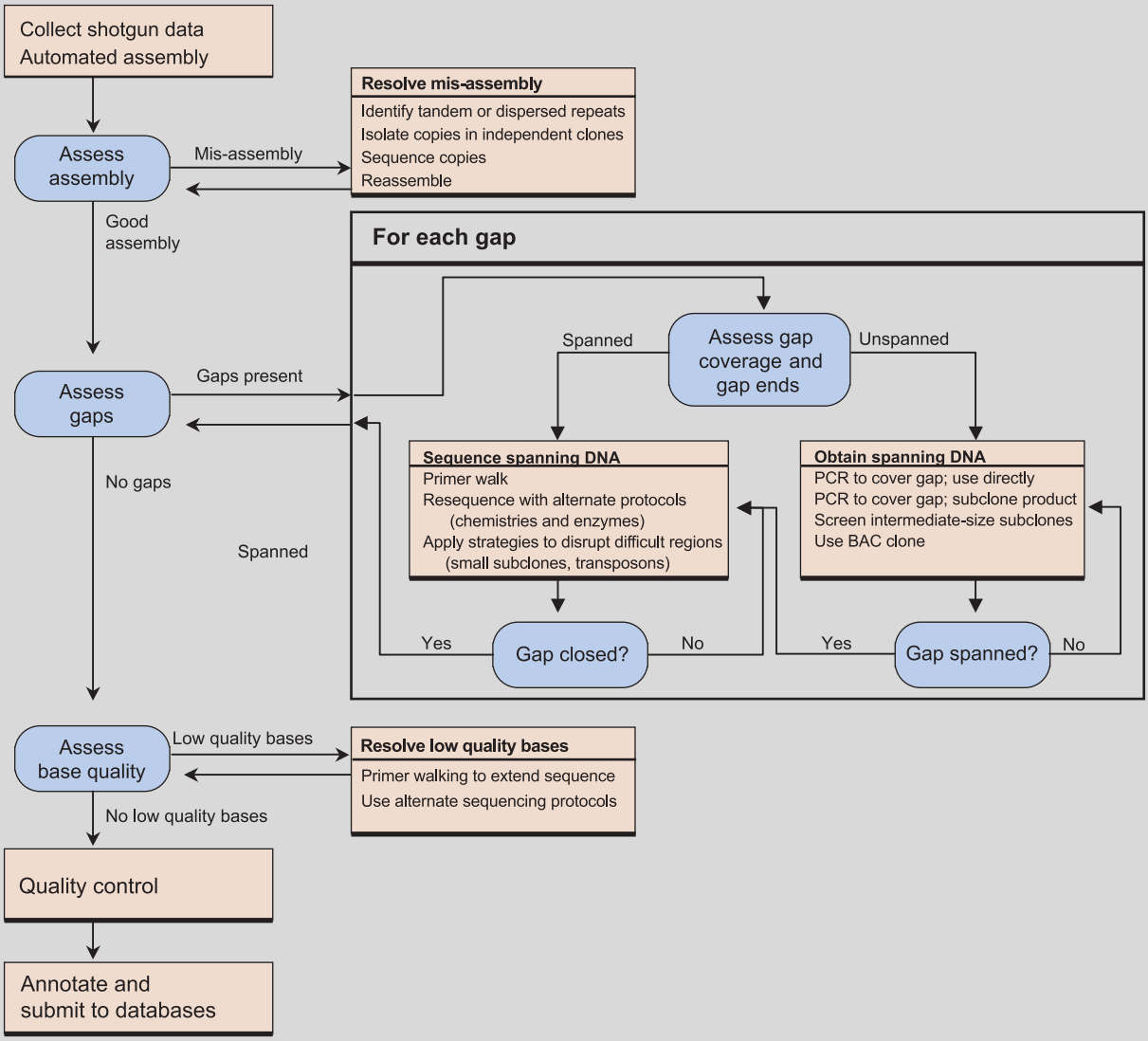
‖Those in the tiling path but for which it has not been possible to obtain finished sequence. Unfinished sequence from these clones is deposited in public databases. These gaps are all listed at 50 kb, reflecting the approximate average size of the gap.

Box 2
Finishing the sequence of clones

Sequencing of large-insert clones began with production of an initial assembly based on shotgun sequence data. For a typical BAC clone, we generated 6–10-fold coverage in paired end-sequences from random, small-insert (2–4 kb) plasmid clones and used computer programs (P. Green, unpublished and see ref. 68) to assemble the data into sequence contigs connected by linking information. Each base was assigned a quality score reflecting its predicted accuracy, based on the underlying shotgun sequence data. The assembly typically had gaps and low-quality regions, with the number varying greatly across clones. These regions are highly enriched in sequences that are difficult to clone or sequence and thus are not represented even after deep (6–10-fold) coverage with random reads. (These regions are also poorly covered by whole-genome shotgun strategies[69].)

The finishing phase converted this draft assembly into a high-quality continuous sequence by obtaining directed information. It involved iterative cycles of computational analysis and laboratory work. Box 2 Fig. 1 shows a simplified flowchart. The first step was to inspect the draft assembly for evidence of mis-assembly, arising from inappropriate merger of repeated sequences. Such evidence would include inconsistent patterns of linking among contigs, regions with unusually high coverage in sequence reads and bases with 'high-quality discrepancies' among the underlying sequence reads. In general, sequence assembly is more straightforward for the clone-based hierarchical shotgun strategy than for the whole-genome shotgun strategy, because the use of clones avoids problems arising from polymorphism and from different copies of repeated regions elsewhere in the genome. Most clones passed assembly inspection, but some failed due to the presence of very similar local dispersed, tandem or inverted repeats. Careful inspection could resolve the problem in some cases, but specific strategies had to be devised in other cases. One approach was to isolate distinct copies of the repeat in subclones of intermediate size (10-kb plasmids or fosmids) and sequence these subclones. Box 2 Fig. 2 illustrates an initially mis-assembled BAC clone from chromosome Y that could be assembled correctly with careful editing.

The second step was gap closure. Because gaps tended to be enriched for problematic sequences, gap closure was challenging; it often required multiple attempts using a variety of alternative methods. Gaps were classified into two types: 'spanned' and 'unspanned'. Spanned gaps were those for which the two flanking contig ends were linked by an end-sequenced plasmid. Most such gaps could be closed by primer-directed sequencing of the plasmid, serially extending the

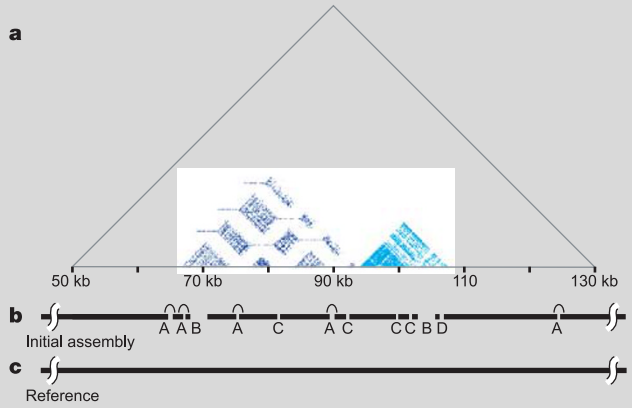


Box 2 Figure 1 Simplified flowchart for finishing of clones.

contig sequence into the gap. Sequence in the gap was often recalcitrant to the standard sequencing protocol (accounting for its absence from the initial shotgun data), making it necessary to use many alternative protocols (different buffers, enzymes and temperature conditions). Some gaps could not be closed by primer walking, because no suitable primer could be found (due to repetitive sequence near the end of the contig) or because sequencing chemistries were unable to penetrate certain secondary structures (such as some inverted repeats). Specialized strategies were used to obtain the missing sequence. For example, problems arising from secondary structure might be overcome by sequencing a small insert library[70] of random subclones with tiny inserts (~100–300 bp, referred to as a "shatter library") or by sequencing from transposon insertions in the plasmid. Unspanned gaps arose where a contig end was not linked to any other contig. It was then necessary to infer adjacency and extend the sequence by other means. Techniques included PCR to other contigs, analysis of various types of subclones from the BAC, and primer walking directly on the BAC[71]. This battery of techniques succeeded in virtually all cases. (In 728 cases, there remains a small region of bases that could not be reliably sequenced; almost all fall in tandem repeat sequences and typically affect tens to hundreds of bases. These cases are annotated in the accessioned clones.)

The third step (which proceeded in parallel to gap closure) was the resolution of low-quality regions. This was accomplished by obtaining additional sequence reads from resequencing of existing shotgun subclones or from primer-directed sequencing.

The final step involved quality control. To confirm the accuracy of the overall assembly, the restriction digestion pattern of the BAC predicted from the finished sequence was compared with the pattern observed experimentally. To confirm accuracy at the nucleotide level, the finished sequence and supporting data were reviewed by human inspection and computational analysis. The finished sequence was then annotated and deposited in public databases.



Box 2 Figure 2 Illustration of a particularly challenging clone. **a**, Central portion of clone RP11-488I11 is illustrated by a triangular dot plot. The base of the triangle represents 80 kb of a 152-kb insert. Each dot represents a perfect match of 20 bases. The region between 65 kb and 94 kb contains four copies of a directly repeated sequence of about 3 kb (horizontal lines), separated by imperfect short tandemly repeated sequence (diamond blocks of dots). The region between 94 kb and 107 kb contains tandemly repeated imperfect copies of a five-base sequence, unrelated to the previous sequence. **b**, Initial assembly of region after completion of shotgun data collection. Two mis-assemblies resulting from the long direct repeats and the absence of all copies (not shown) were resolved by manual editing, after which 12 gaps remained in the clone. Five of these (labelled A) were spanned by plasmid subclones and were closed by primer walking. Two gaps (labelled B) were larger; after initial walks failed, these gaps were closed by sequencing short insert libraries prepared from PCR products. Four other gaps (labelled C) were not spanned by plasmid clones but were closed by primer walks on PCR products. One gap (labelled D) was closed by primer walks and extensive manual editing. **c**, The finished clone with all gaps closed.

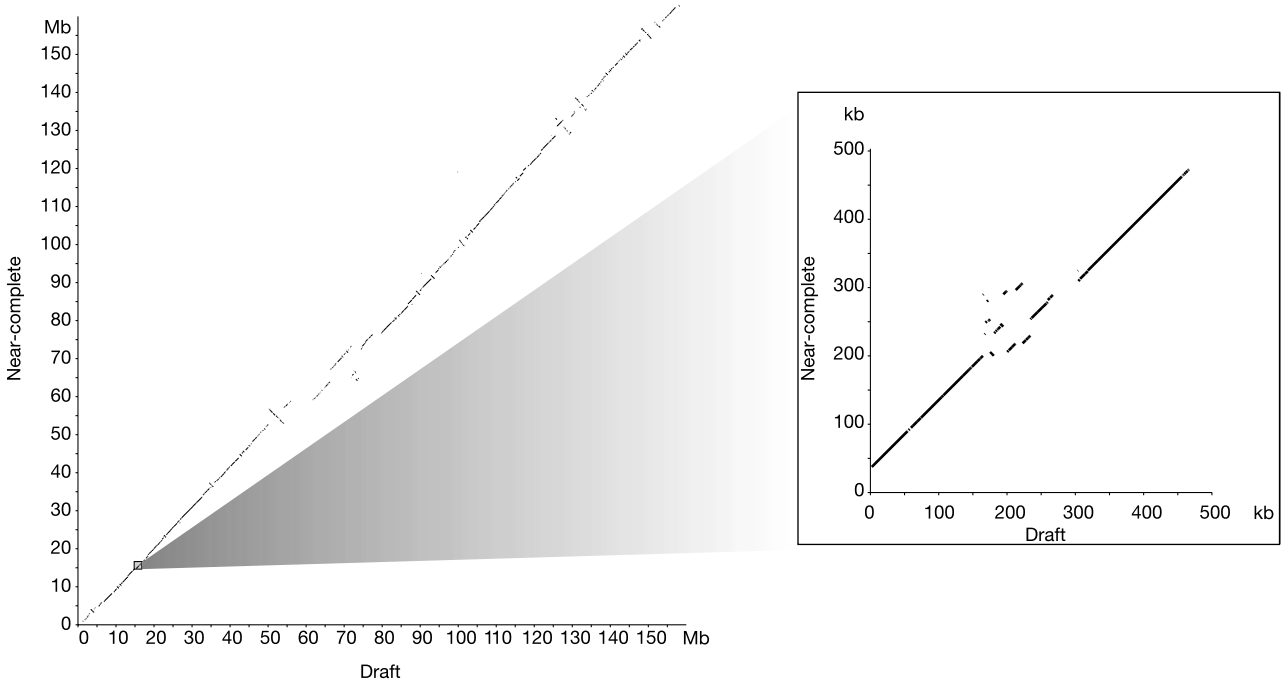


Figure 1 Comparison of previous draft sequence with current near-complete sequence of chromosome 7 (ref. 24). At large scale, there was good collinearity between draft and near-complete sequence, although some inversions were present in the draft due to lack of sufficient anchors in some regions. At finer scale, the draft sequence contained some sequence contigs for which order and orientation were not known. The inset shows a region of 500 kb with sequence derived from three overlapping BACs. BACs at each end were finished at the time of draft assembly, whereas the middle BAC was at an early stage of shotgun coverage in which contigs were not yet ordered and oriented.

permanent foundation for biomedical research, it was important to assess its quality and to characterize its remaining defects. For this purpose, we used a number of comparisons and consistency checks.

Assessment of accuracy

Tests of accuracy were designed to detect potential problems that may have occurred in clone-based sequencing. This may include errors in assembling the finished sequence within individual clones, and errors in concatenating adjacent finished clones to create the final product. The analysis was complicated by the presence of polymorphism in the human population, because differences between sequence clones may reflect either errors or polymorphism.

Independent quality assessment. Quality assessment (QA) exercises were performed regularly throughout the HGP[31]. In the final stages, an independent group examined a random sample of finished clones by generating additional data and generating new assemblies[32]. Briefly, this QA analysis examined ~34 Mb and found an error rate of 1.1 per 100 kb for small events (≤50 bp, with average size of 1.3 bp) and 0.03 per 100 kb for large events (>50 bp). The small events consisted largely of single-base substitutions, whereas the remaining small and large events primarily concerned the number of consecutive copies of a tandem repeat[32].

Table 3 **Chromosome arm length and contiguity in draft and reference sequence**

Chromosome	Euch. length* (bp)	N50† draft§ (bp)	Build 35 N50 ref‖ (bp)	Build 35 N-average ref§ (bp)
1p	121,147,476	81,895	16,783,271	33,566,574
1q	104,135,370	45,843	56,331,646	36,675,159
2p	91,748,045	68,853	68,373,980	53,478,029
2q	148,270,183	50,481	84,213,156	54,482,973
3p	90,587,544	39,322	66,080,833	54,853,737
3q	106,018,194	35,734	100,530,261	96,935,077
4p	49,501,045	36,494	9,040,907	13,797,821
4q	138,910,172	31,876	92,070,735	66,386,026
5p	46,441,398	59,470	46,378,398	46,378,398
5q	131,416,467	81,416	41,199,371	33,564,217
6p	58,938,125	251,648	48,945,890	42,200,138
6q	109,037,573	150,424	61,695,806	46,408,435
7p	57,864,988	399,235	47,497,097	40,050,874
7q	97,763,150	298,612	64,426,257	46,810,648
8p	43,958,052	40,151	9,464,880	9,872,060
8q	99,316,773	37,528	57,155,273	47,945,192
9p	46,035,928	87,767	39,435,726	34,619,306
9q	74,393,339	43,983	40,394,264	29,078,785
10p	39,244,941	48,121	20,794,160	15,791,760
10q	93,788,686	47,401	30,112,613	31,833,318
11p	51,450,781	34,383	49,571,094	48,044,101
11q	80,001,602	42,527	17,911,127	26,070,918
12p	34,747,961	197,985	27,615,668	23,435,010
12q	96,306,849	47,272	32,815,934	29,605,325
13p	acro arm	n/a	n/a	n/a
13q	96,274,979	70,497	67,740,325	54,830,719
14p	acro arm	n/a	n/a	n/a
14q	88,298,584	1,370,997	88,290,585	88,290,585
15p	acro arm	n/a	n/a	n/a
15q	82,078,915	30,303	53,619,965	38,049,097
16p	35,143,302	160,390	25,336,229	20,462,803
16q	43,883,952	86,933	42,003,582	40,305,188
17p	22,187,133	114,901	21,163,833	20,341,190
17q	56,487,608	82,866	11,472,733	15,591,618
18p	15,400,898	59,951	15,400,898	15,400,898
18q	59,352,257	50,087	33,548,238	26,073,241
19p	26,923,622	82,369	15,825,424	12,506,733
19q	33,888,028	167,408	31,383,029	31,383,029
20p	26,267,569	1,436,102	26,259,569	26,259,569
20q	34,402,734	1,301,134	26,144,333	21,428,992
21p¶	490,223	n/a	490,223	490,223
21q	33,684,323	28,515,322	28,617,429	24,743,931
22p	acro arm	n/a	n/a	n/a
22q	35,224,709	23,048,103	23,276,302	16,327,958
Xp	58,465,033	173,718	33,063,353	22,383,515
Xq	93,359,231	277,548	27,718,692	25,766,623
Yp	11,237,315	5,778,849	6,265,435	4,331,076
Yq	15,464,376	1,026,317	10,002,238	8,061,778
All arms	2,879,539,433	82,663	38,509,590	40,970,092

*Chromosome arm lengths refer to estimated length of euchromatic portions of each arm.
†N50 denotes the contig length *x* (for a chromosome arm or entire genome) such that half of all nucleotides reside in contigs of length at least *x*.
‡'N50 draft' reports this number for the draft sequence[15].
§The value for the near-complete reference sequence reported here.
‖Average contig length in the near-complete sequence for a randomly chosen nucleotide (or, equivalently, average length contigs weighted by length).
¶Chromosome 21p is an exception to the generalization that the acrocentric arms only contain heterochromatin—there is a 281-kb contig within chr 21p11.2.

Analysis of clone overlap. We extended the QA analysis to a larger region (~174 Mb), by examining overlapping sequence between consecutive finished large-insert clones. If two such clones derive from the same copy of the human genome, any sequence differences in the overlap must reflect an error in one of the two clones. By comparing independent clones, this quality assessment method also has the ability to detect cloning artefacts. We examined 4,356 substantially overlapping clones derived from the same library; half are expected to be derived from the same haplotype and half from a different haplotype. We counted the number of single-base mismatches (ignoring insertion/deletions (indels)) in the overlapping regions. The resulting distribution (Fig. 2a) is bimodal. The first peak is consistent with expectation for clones from the same haplotype, with a sequencing error rate of $\sim 10^{-5}$ per bp. The second peak is consistent with the expectation for clones from different haplotypes, with a polymorphism rate of $\sim 10^{-3}$ per bp; this peak matches the distribution seen for clones from different libraries.

We then examined overlapping clones likely to be from the same haplotype (with no single-base mismatches) and counted the discrepancy rate for indels (Fig. 2b). The error rate (estimated as half the discrepancy rate) is ~0.55 events per 100 kb, with the vast majority being in tandem repeats. By contrast, clones from different libraries show a discrepancy rate that is at least 20-fold higher. Overall, the analysis indicates that the overall error rate (reflecting both sequence error and cloning artefacts) is 20–100-fold lower than the human polymorphism rate.

Analysis of junctions. We assessed longer-range integrity of the genome sequence by studying read pairs from large insert clones. Specifically, we created a fosmid library carrying randomly sheared human DNA and sequenced both ends of the insert of ~750,000 clones. Fosmid clones are particularly useful because their insert sizes cluster tightly around 40 kb, due to packaging constraints. We aligned the fosmid end sequences to the genome sequence. Both ends could be mapped to unique locations in the human genome in most cases (86%), and these two locations were within 39.5 ± 7.5 kb in 99% of cases. Some fosmids could not be uniquely placed because one or both ends consisted almost entirely of repeat sequence. Using the uniquely placed fosmids (which provide about eightfold clone coverage of the euchromatic genome), we sought to obtain independent confirmation of the order, orientation and adjacency of the junction between consecutive finished large-insert clones used to construct the genome sequence. The junction was considered 'supported' if spanned by one or more consistently placed fosmids. In all, ~97% of junctions were supported. About half of the remaining junctions were supported by fosmids with unique placement at one end but multiple placements at the other end. Overall, the analysis provided strong support for accuracy of the junctions underlying the current genome sequence.

Search for deletions. We next scanned the genome sequence for evidence of deletions of several kilobases in size, using the same fosmid data set. At each point, we calculated the 'apparent size' of each fosmid spanning the point (defined as the distance between the location of the end sequences in the current genome sequence) and then calculated the 'average apparent size' for all the fosmids spanning the point. We searched for regions where the observed size fell far below expectation (<3.5 standard deviations (s.d.)), suggesting a large difference between the genome sequence and the source DNA for the fosmid library (Fig. 3). Such differences could reflect either an error in the genome sequence, a deletion in the fosmid clone, or a deletion polymorphism between the DNA sources. (Given the number of fosmids used, this analysis has ~50% sensitivity to detect deletions of 3–30 kb. Because the

methodology cannot detect deletions larger than a fosmid, we also analysed discrepant fosmid links, which could reflect deletions. See Methods in Supplementary Information.)

We found 242 candidate regions, with suggestive evidence for deletions (average apparent size ~5 kb). These regions were then scrutinized by alignment with the recently obtained draft sequence of the chimpanzee genome (R. H. Waterston, personal communication). Because the human and chimp genomes align with relatively few large indels (indels >2 kb occur at ~1 per 100 kb), this comparison should highlight true deletions. The chimpanzee comparison supported the presence of deletions in 35% of cases. A subset of these was then tested by polymerase chain reaction (PCR) analysis of genomic DNA from multiple individuals. Roughly two-thirds appear to represent polymorphic deletions in the human population and one-third represent actual errors in the current genome sequence. Overall, the results indicate that the current genome sequence is likely to contain perhaps 50–100 erroneous deletions (average size ~5 kb), which could be due to assembly errors or mutations occurring during propagation of large insert clones. Analysis of a larger collection of fosmids could probably pinpoint the majority of these errors, allowing them to be corrected.

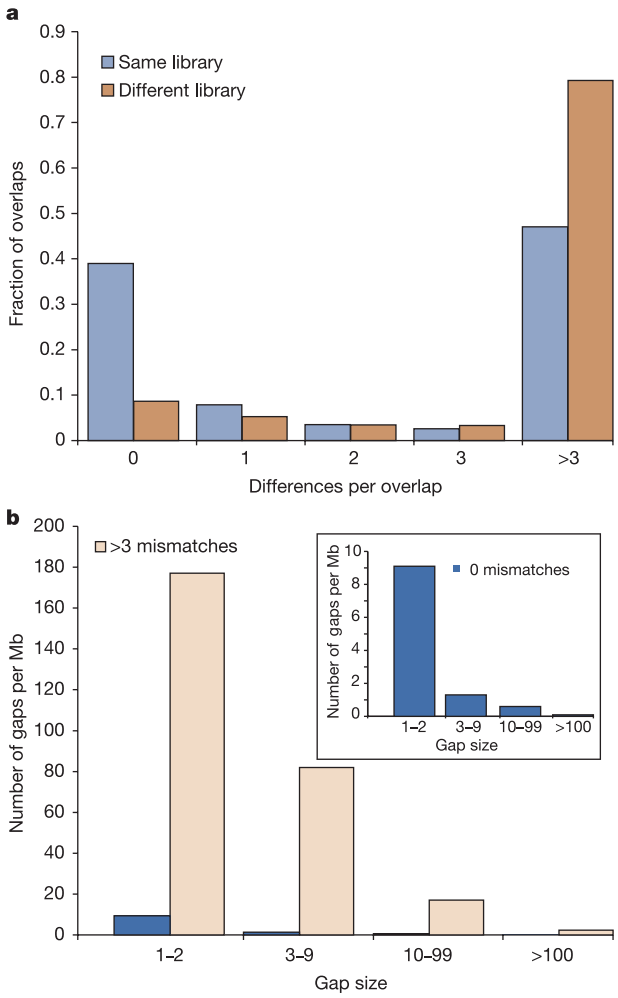


Figure 2 Assessment of potential errors by analysis of BAC overlaps. **a**, Single-base differences between overlapping finished BAC clones (with ≥5 kb overlap). The number of single-base differences in overlaps for clones from the same library and from different libraries is plotted. The results are consistent with half of the clones from the same library representing identical underlying DNA sequence with low error rate, and half representing different haplotypes as expected. **b**, Insertion/deletion (indel) differences between overlapping clones. The number of indels per Mb for a given size range is compared for clones with no single-base mismatches (presumed to be derived from the same haploid source) and >3 single-base mismatches (presumed to be derived from different haploid sources). Indels in the former class primarily represent errors in finished sequence; they occur at ~20-fold lower frequency (inset) than indels in the latter class, which primarily represent polymorphic differences.

Assessment of coverage

Tests of coverage were designed to measure the proportion of the euchromatic genome missing from the current genome sequence, by assessing the presence of independently sampled human sequences such as complementary DNA clones and random genomic clones.

Analysis of cDNAs. We tested for the presence of known cDNA sequences from public databases (REFSEQ[33] and MGC[34]). The analysis[35] involved 17,458 distinct gene loci spanning 925 Mb of genomic sequence. The vast majority (99.74%) could be confidently aligned to the current genome sequence over virtually their complete length with high sequence identity (at a level consistent with the expected polymorphism rate and the performance of the alignment program). A few of these (0.5%) showed strong alignment to more than one locus. A few others (0.04%) showed unusually high sequence difference (>2%), but these were nearly all immunologically related genes (such as major histocompatibility loci and immunoglobulin-related loci) known to be highly polymorphic.

We examined the remaining cases (0.28%). The cDNA sequence appeared to be completely absent in 0.06% of cases and partially absent, with a contiguous segment missing, in 0.23% of cases. For

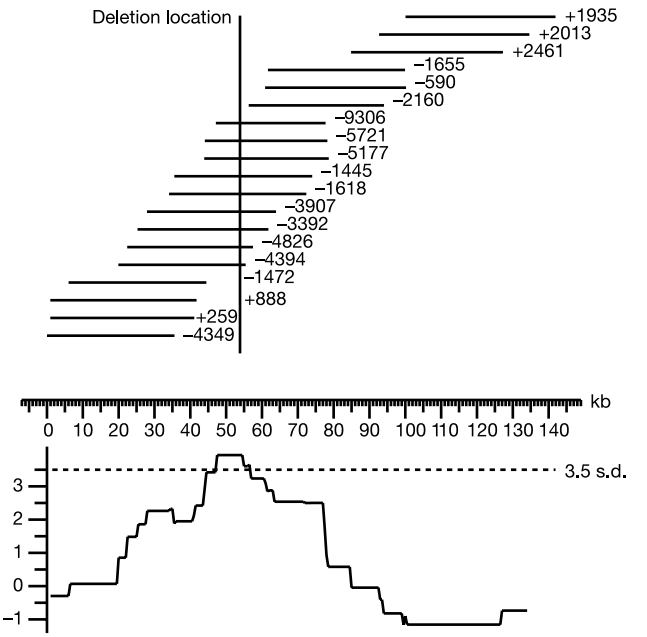


Figure 3 Detection of potential insertions or deletions using paired-end fosmid reads. The top portion shows fosmids along a region of chromosome 10 (centred at nucleotide 46,915,451), mapped by virtue of their paired-end sequences. The difference between inferred length, calculated from the location of fosmid ends in finished sequence, and average length for the entire library, is shown to the right of each clone. For each point, the standard deviation of the local average difference for all spanning fosmids is plotted below; the threshold of 3.5 standard deviations is indicated by a dotted line. The region from 45 to 55 kb is inferred to contain a length difference between the fosmids and finished sequence. Comparison with available chimpanzee sequence further localized the difference (vertical line). Experimental analysis (PCR from clones used for finished sequence and the fosmid library, as well as from 24 random humans) confirmed the difference, and showed that it is due to an insertion/deletion polymorphism of 5.8 kb. The majority of length differences detected by this analysis appear to represent polymorphisms, not sequence errors.

almost all of completely absent cDNAs, the genomic location of the gene was known or could be inferred and corresponds to a gap in the current genome sequence. For the partially absent cDNAs, more than half of the cases lie adjacent to gaps. The remainder may represent either errors in the current genome sequence or polymorphic deletions; these are being investigated further. Overall, the proportion of cDNA sequence that is missing from the genome sequence is only 0.08% of the total. This may underestimate the proportion of genome missing from the finished sequence, however, because focused efforts were made to capture genomic sequence containing missing messenger RNAs.

Analysis of random genomic plasmids. As an additional and broader test of coverage, we analysed paired end-sequences from 5,000 small-insert (3–4 kb) plasmids generated as part of a human single nucleotide polymorphism (SNP) discovery project (see Methods). After excluding heterochromatic repeats and other artefacts, we found that 99.3% of the reads could be reliably aligned to the finished sequence. For 0.6% of the reads, neither end could be aligned; these probably lie in known gaps. For another 0.1% of the reads, exactly one end could be placed; some fell next to known gaps, whereas others appear to represent indel differences between the reference sequence and the source DNA for the plasmid library. The overall analysis indicates that <1% of the euchromatic genome is missing from the finished sequence. Together, the cDNA and plasmid analyses indicate that the current genome sequence contains more than 99% of the euchromatic portion of the human genome.

Characterization of remaining gaps

The current genome sequence contains 341 gaps, which could not

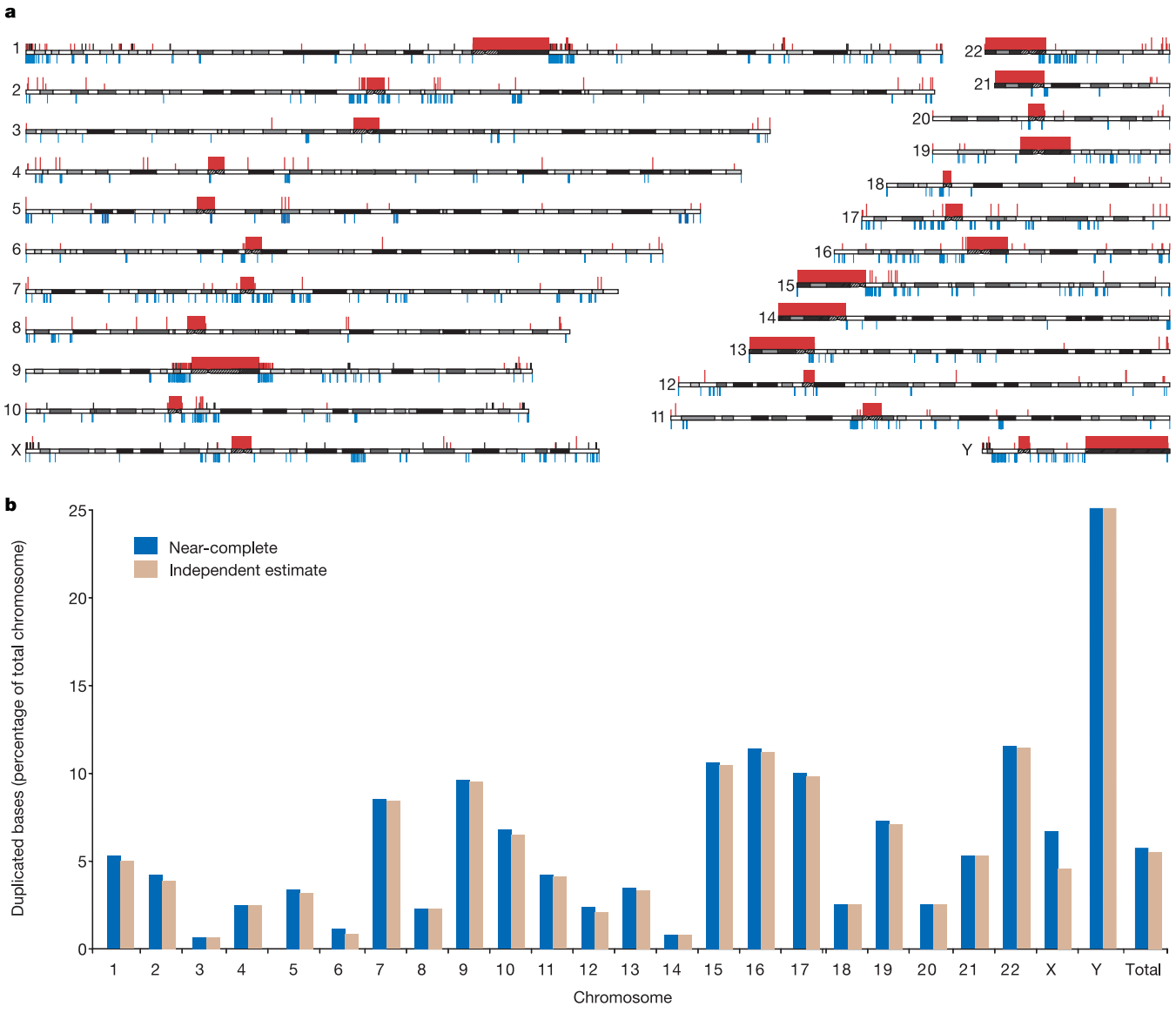


Figure 4 Segmental duplications across the genome. **a**, Segmental duplications and sequence gaps across the genome. Segmental duplications are indicated below the chromosomes in blue (length ≥10 kb and sequence identity ≥95%). Large duplications are shown to approximate scale; smaller ones are indicated as ticks. Sequence gaps are indicated above the chromosomes in red. Large gaps (>300 kb) are shown to approximate scale; smaller gaps are indicated as ticks with those that are 50 kb or smaller shown as shorter ticks. Unfinished clones are indicated as black ticks. **b**, Percentage of large segmental duplications by chromosome. This count includes both interchromosomal and intrachromosomal duplications with length ≥1 kb and sequence identity ≥90%. The blue bars show the result of direct analysis of near-complete sequence. The gold bars show an independent estimate[65] using whole-genome shotgun data to correct for potential mis-assembly of such segmental duplications. The strong agreement suggests that most segmental duplications are properly represented in near-complete genome sequence. The discrepancy for chromosome X is probably a result of errors in the independent estimate, due to limited coverage and diversity of data from this chromosome[15].

be closed with available techniques. We briefly describe the nature of these gaps and discuss the prospects for eventual closure. (See Supplementary Information Notes 2 and 4.)

Heterochromatic regions (33 gaps). The heterochromatic regions of the human genome were not targeted by the HGP, because their highly repetitive properties make them largely refractory to current cloning and sequencing strategies. There are 33 heterochromatic regions falling into four types. The 24 centromeres (~50 Mb) consist largely of alpha satellite repeats, of which ~15 types exist; these monomeric repeats are arranged into higher-order arrays

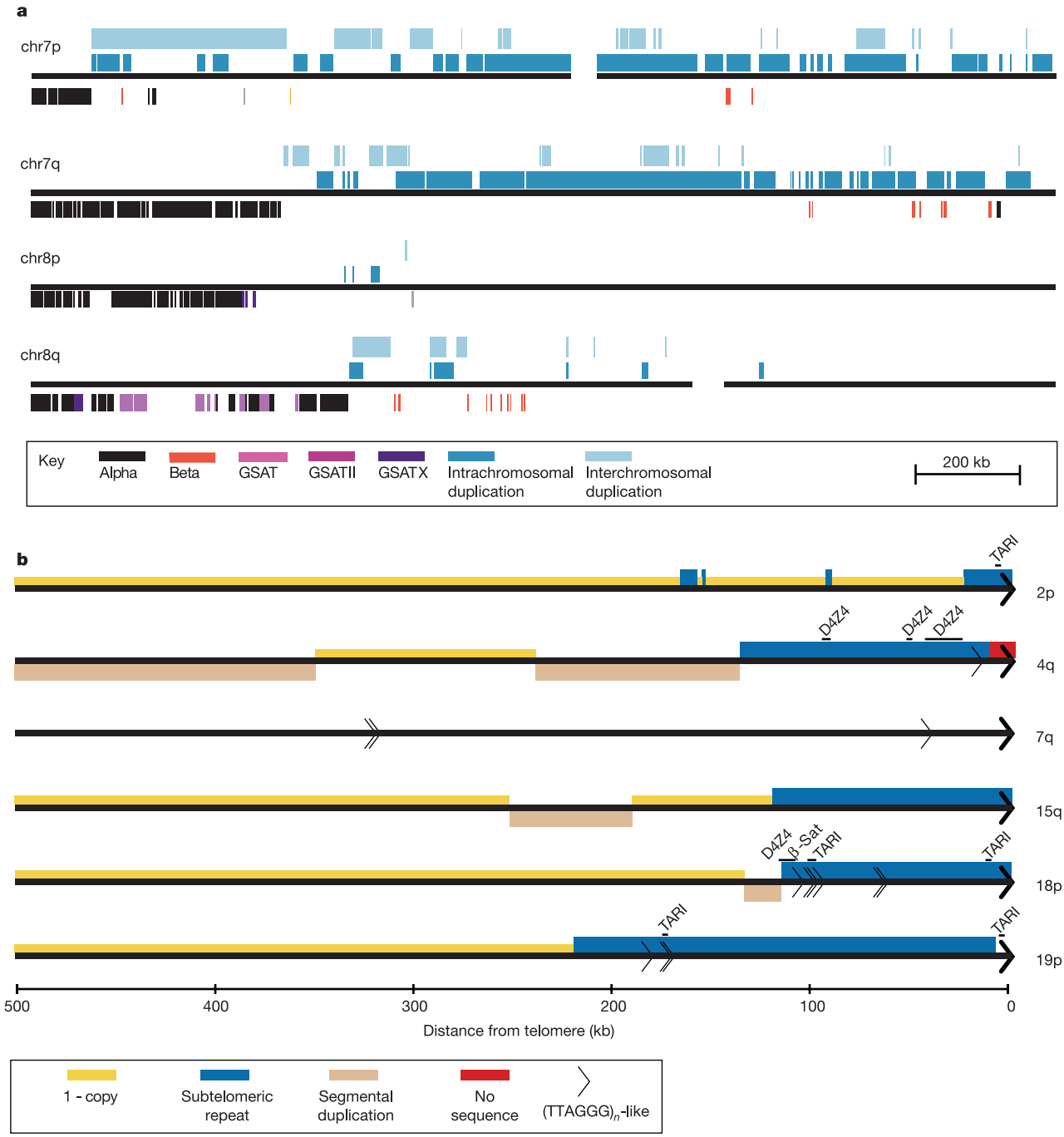


Figure 5 Examples of repeat structure near centromeres and telomeres. **a**, Repeats in pericentric regions of chromosomes 7 and 8. The most proximal regions are crowded with alpha satellite sequences and other centromeric repeats; composition, density and order may vary considerably between chromosome arms[62]. Just outside this region, there is usually a high density of inter- and intra-chromosomal duplication. For details, see text and refs 39, 40, 66 and 67. **b**, Sequence organization in human subtelomeric DNA regions. The terminal repeat tract consists of 2–15 kb of simple repeat sequence $(TTAGGG)_n$ and is indicated by the black arrow at right. Short (50–250 bp) and often degenerate (TTAGGG) tracts (internal black arrows) are highly enriched (>25-fold) in subtelomeric DNA relative to elsewhere in the genome. A subtelomeric repeat (Srpt) region (blue) consists of a mosaic patchwork of segmentally duplicated DNA tracts that occur in two or more subtelomere regions and range in size from <10 kb to >300 kb. TAR1, D4Z4 and beta satellite sequences are frequently associated with Srpt regions. Proximal to the Srpt region is chromosome-specific genomic DNA, typically with a high GC content and high gene density. Stretches of segmentally duplicated DNA that occur only once within subtelomeric regions (tan) are interspersed with 1-copy subtelomeric DNA (yellow) in a telomere-specific fashion. Overall, segmentally duplicated DNA comprises approximately 25% of the most telomeric 500 kb of the chromosome, a fivefold enrichment over the genome-wide average.

distinct to specific chromosomes, which are tandemly repeated with slight sequence variations. The three secondary constrictions are immediately adjacent to the centromere on chromosome arms 1q, 9q and 16q and contain various satellite repeats (beta, gamma, satellite I, II, III). The five acrocentric chromosome arms 13p, 14p, 15p, 21p and 22p encode the 5S, 18S and 28S ribosomal RNA genes, which lie on a 43-kb sequence present in ~50 tandem copies on each arm and are flanked by additional repeats arranged in complex structures. Finally, there is a single large region on distal Yq composed primarily of thousands of copies of several repeat families. The heterochromatic regions all tend to be highly polymorphic in length in the human population.

Euchromatic boundary regions (35 gaps). The euchromatic regions of the human genome are bounded proximally by heterochromatin and distally by a telomere consisting of several kilobases of the hexamer repeat TTAGGG. We examined the current genome sequence for evidence of the expected boundaries on the 43 euchromatic arms. (See Supplementary Information Note 4.) At the proximal ends, 30 of the 43 cases show sequence characteristic of either heterochromatin or immediately flanking regions (such as higher-order centromeric repeats, stretches of at least 10 kb of monomeric alpha satellite repeat or other pericentromeric repeats). We cannot exclude the possibility that there is additional unique sequence between this point and the proximal heterochromatin; but efforts to extend the finished sequence further were unsuccessful. In the remaining 13 cases, the finished sequence contains no evidence of heterochromatin-related sequence. At the telomeric ends, 21 of the 43 cases show continuous sequence extending to the telomeric repeat. This sequence was typically obtained by isolation and sequencing of half-YAC clones spanning to the telomere[36]. An additional 18 cases are sequence gaps, in which half-YACs reaching to the telomere were isolated but finished sequence could not be obtained. The remaining four cases are physical gaps, in which large-insert clones extending to the telomere could not be obtained.

Euchromatic interior regions (273 gaps). The remaining gaps are located within the current genome sequence. These consist of 215 physical gaps for which no clones could be isolated, and 58 sequence gaps for which clones were found but reliable finished sequence could not be obtained. The physical gaps are greatly enriched in regions of segmental duplication (Fig. 4a). Roughly half of these gaps (52%) are flanked by segmental duplications with >90% sequence identity, although such duplications comprise only ~5.3% of the euchromatic genome (Fig. 4b). Such segmental duplications are especially frequent in pericentromeric regions, and gaps are notably more frequent in these regions. The association of gaps with segmental duplications is examined in detail elsewhere[37].

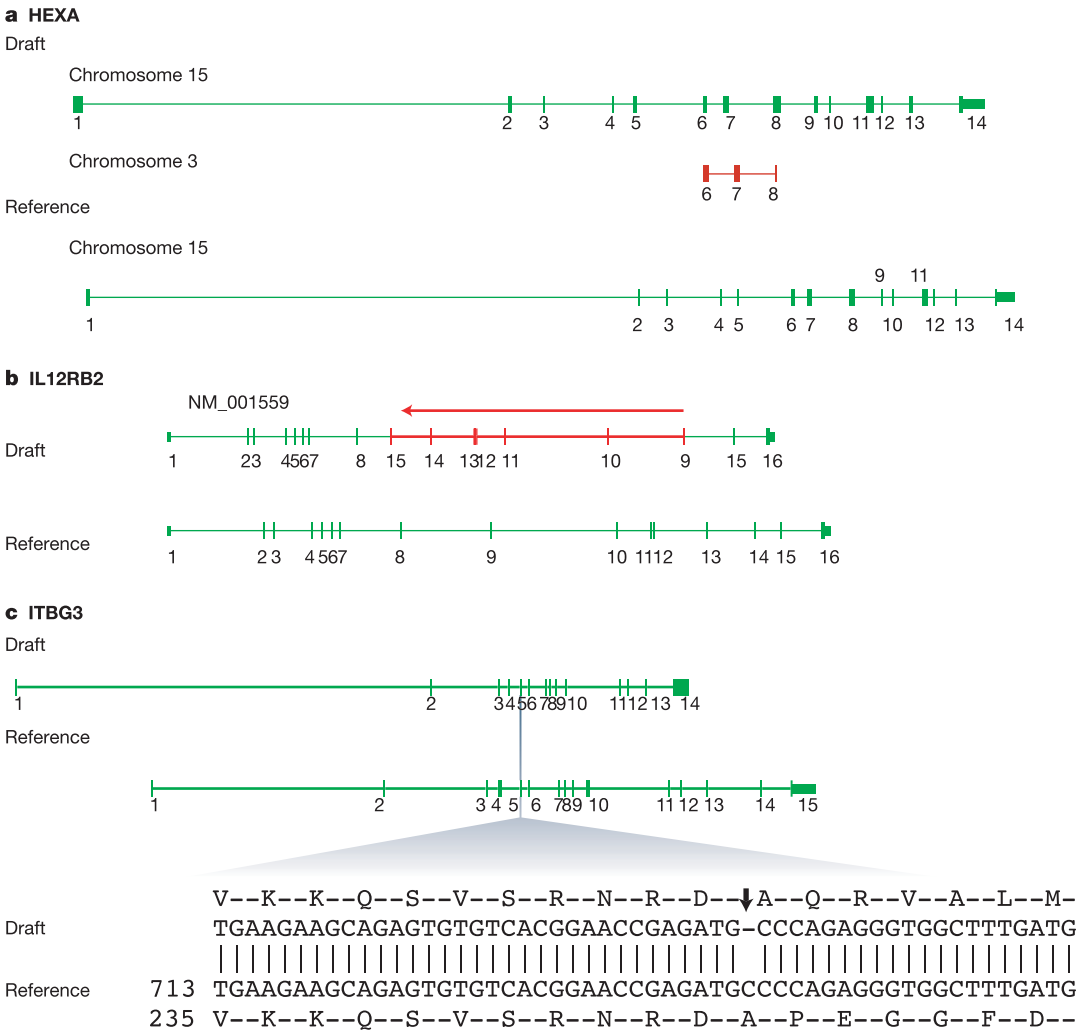


Figure 6 Examples of genes corrected in the near-complete sequence. **a**, In draft sequence, exons 6,7 and 8 of hexosaminidase A (HEXA) gene were present on chromosome 3 in addition to their correct location on chromosome 15. **b**, In draft sequence, interleukin 12 beta-2 receptor (IL12RB2) contained an inversion of several internal exons and a duplication of exon 15. **c**, In draft sequence, a single-base deletion in exon 5 of integrin beta-3 precursor (ITGB3) caused a frameshift at amino acid 247. The terminal exon was also missing.

The most extreme case occurs near the centromere of chromosome 9. The most proximal 5 Mb on 9p and 4 Mb on 9q comprise a mere 0.3% of the genome, but account for ~12% of the physical gaps in the euchromatic sequence. These two pericentric regions are unique in the genome with respect to density of segmental duplication and the average degree of intrachromosomal sequence identity (98.7%), and the two regions have many highly similar sequences in common. The high sequence similarity between the two regions is likely to be the reason for a polymorphic inversion of the centric heterochromatin on chromosome 9, present at a frequency of ~1% in the human population[28]. Other proximal regions also show a higher-than-average density of gaps. For example, the proximal 2 Mb on the remaining 41 euchromatic arms comprise 2.9% of the genome but harbour 13.3% of the gaps. Nearly all of these proximal gaps are flanked by segmental duplications (Fig. 5a). There is also a clustering of such gaps in subtelomeric regions. The terminal 1 Mb on the 43 euchromatic arms represents 1.5% of the genome, but contains ~14% of the total gaps; nearly all of these gaps are also flanked by segmental duplications (Fig. 5b).

Closing the remaining gaps. Although the euchromatic genome sequence has reached a much higher degree of completion than had been anticipated, it still remains incomplete with ~1% of the euchromatin residing in 308 gaps. These represent regions that could not be reliably mapped, cloned and sequenced with current methods. Rather than applying further brute force, it is now time to develop focused strategies to resolve the regions.

The remaining euchromatic gaps probably reflect two major issues. The first pertains to regions harbouring segmentally duplicated sequence. Such regions are challenging to map because it can be extremely difficult to discern whether two clones with small sequence differences represent different loci or different alleles at a single locus. This challenge was eventually resolved for chromosome Y (ref. 23) (which is especially rich in segmental duplication) by exploiting the fact that the chromosome is haploid in males. By using DNA from a single haploid source, it was possible to rely on differences at only a handful of nucleotides to distinguish repeated sequences. This approach could be applied to the rest of the genome by using appropriate haploid sources, such as a hydatidiform mole or monochromosomal hybrids. (In both instances, use of parental controls to guard against being misled by somatic rearrangements would be well advised.) It may be useful to test these approaches on individual chromosomes. The second issue is that some gaps are likely to correspond to regions that cannot be efficiently propagated in current large-insert vectors and hosts. It may be useful to test new kinds of large-insert libraries for clones containing unique sequences not contained in the current human genome sequence (perhaps seeded by probes derived from random small-insert genomic plasmids, as discussed above). In addition, genome completion may benefit from long-range mapping techniques such as optical mapping[38], which may provide independent information about difficult regions.

Completing the euchromatic sequence is an important goal, but is clearly now a research effort rather than a high-throughput project. Sequencing the human heterochromatin poses an even greater challenge. The current human sequence penetrates only the periphery of the heterochromatin—for example, the pericentric regions on a few chromosome arms[39,40]. This progress has required concerted efforts with specialized mapping techniques and painstaking assembly. The fundamental issue is that current shotgun

Table 4 **Human paralogous genes**

Chromosome	Cluster size in human genome	Minimum size in ancestral genome	Genes involved in recent duplications	Gene family
2	30	8	26	Immunoglobulin K chain V
11	64	50	23	Olfactory receptor
19	23	5	19	KRAB zinc-finger protein
14	21	9	15	Immunoglobulin heavy chain
9	16	4	13	Interferon α
1	34	25	13	Olfactory receptor
19	18	9	13	Leukocyte and NK cell immunoglobulin-like receptors
22	20	11	12	Immunoglobulin λ chain V-region
1	13	3	11	PRAME/MAPE family (cancer/germ line antigen)
16	11	2	11	Immunoglobulin heavy chain
19	10	1	10	Pregnancy-specific β-1-glycoprotein
11	59	54	10	Olfactory receptor
X	9	1	9	Hypothetical gene LOC255313 expressed in testis
17	13	9	8	Olfactory receptor
4	9	3	7	UDP-glucuronosyltransferase; steroid metabolism
12	14	8	7	Taste receptor, type 2
19	16	10	7	FDZF2-like KRAB zinc-finger protein
X	7	1	7	SSX-like KRAB zinc-finger protein (CT antigens)
X	8	2	7	MAGE (CT antigens)
Y	7	1	7	Testis-specific Y-encoded (TSPY) protein
4	9	5	7	CXCL1/MIP2-like small chemokines
7	6	1	6	Postmeiotic segregation increased-2 (DNA mismatch repair)
8	6	1	6	FLJ00326 hypothetical protein
11	6	1	6	TRIM48, testis-specific RING finger protein
16	6	1	6	Tumour protein p53-inducible gene, TP53TG3
Y	6	1	6	Testis-specific Y-encoded (TSPY) protein
6	7	3	6	Butyrophilin subfamilies 2 and 3
11	23	19	6	Olfactory receptor
19	18	14	6	Gonadotropin-inducible transcription repressor-2-like
7	5	1	5	Williams Beuren syndrome chromosome region 19 protein
8	5	1	5	Exonuclease GOR
11	5	1	5	Yeast Ssu72p-like protein
14	6	2	5	Immunoglobulin α, δ, γ chains
16	5	1	5	Metallothionein 2A-like
17	5	1	5	Growth hormone gene cluster
19	5	1	5	Testis-specific transcriptional repressor
19	5	1	5	Choriogonadotropin beta (placental hormones)
X	6	2	5	GAGE (CT antigens)
X	5	1	5	XAGE (CT antigens)
X	5	1	5	Sarcoma antigen SAGE (CT antigens)
X	5	1	5	SPAN-X; sperm protein (CT antigens)

The above table includes clusters of human paralogous genes having at least five genes involved in recent gene duplications. Recent is defined as divergence $K_S \leq 0.3$ (see text), corresponding roughly to the divergence of primate and rodent lineage. Ancestral cluster size is the minimum required to account for existing clusters in human, assuming no gene losses in the human-specific lineage.

strategies are poorly suited to assembling large, highly repetitive regions. The hierarchical shotgun strategy faces the challenge of accurate assembly of individual BACs and accurate overlap of BAC clones, with the underlying data consisting of nearly identical sequence; the whole-genome shotgun strategy compounds these problems. Conceivably, the hierarchical strategy could be adapted as was done for repetitive regions of chromosome Y. Approaches might include the use of the following: haploid DNA sources to restrict the problem to a single haplotype; single chromosome sources to avoid confusion among related centromeres on different chromosomes; sheared BAC libraries to avoid biases caused by the unusual distribution of restriction sites within the repeat sequences; assembly based on rare base differences that distinguish near-identical repeats; cloning vectors that minimize rearrangements; and subclone libraries of varying insert lengths. Such an approach will also require ensuring accurate recovery and stability of heterochromatic regions in large-insert clones. Even so, the path is likely to be arduous and expensive to obtain regions of uncertain information content. Alternatively, it may be possible to develop new approaches. These might include methods to obtain much longer effective read lengths, directed reads from known locations and long-range mapping information about the location of rare base differences among repeat copies (such as optical mapping[38] or padlock probes[41]).

Examples of utility of near-complete sequence

The present genome sequence enables far more precise analyses of the human genome, especially those that depend sensitively on high accuracy and near-completeness. Rather than revisit all of the analyses in our initial analysis of the human genome, we have chosen four examples that illustrate the utility of the current near-complete sequence.

Segmental duplications

The human genome is notable for its high proportion of recent segmental duplications. They are of great medical interest because their unusual structure often predisposes them to deletion or rearrangement with consequent phenotypic effects; prominent examples include the Williams syndrome region (7q), Charcot–Marie–Tooth region (17p), DiGeorge syndrome region (22q) and the AZF-C region (Y)[42]. Some regions of segmental duplication have also recently been shown to be evolutionary nurseries in which coding sequences are undergoing strong positive selection[43]. Accurate analysis of segmental duplications was previously impossible because the draft sequence also contained a high degree of artefactual duplication. This difficulty was recognized at the time and the approximate proportion of true and artefactual duplication was inferred indirectly. With near-complete sequence, the artefacts are now largely eliminated and true segmental duplications can be reliably studied.

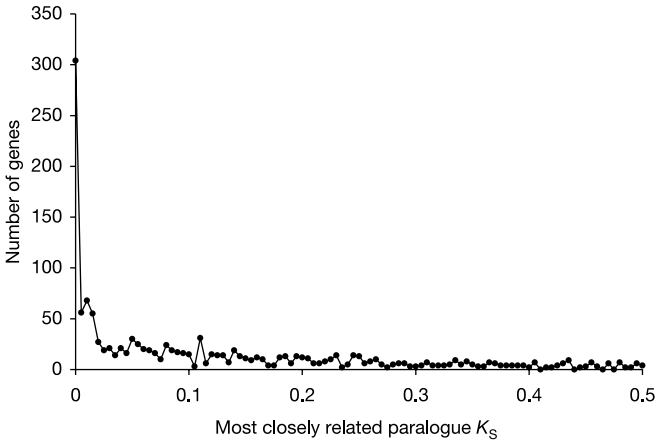


Figure 7 Distribution of K_S for recent gene duplications. K_S, the number of synonymous substitutions per synonymous site, was estimated for each gene from alignment with its most closely related human paralogue. This provides an indication of evolutionary time since divergence.

On the basis of the current sequence, segmental duplications cover ~5.3% of the euchromatic genome. (Here, segmental duplications are counted as regions that are not transposable element copies, are ≥1 kb in length and have sequence identity ≥90%; this corresponds to duplication within the past ~40 million years.) The proportion of segmental duplication and the degree of sequence identity are clearly substantially higher in the human genome than in the mouse[44] or rat[45] genomes (although precise figures for the rodent genomes must await finished sequence). The use of large insert clones, representing a single haplotype, was critical in resolving these regions. The distribution of segmental duplication varies widely across chromosomes, as does the proportion of intrachromosomal versus interchromosomal duplications[15] (Fig 4b). The most extreme case is chromosome Y, which carries segmental duplication along >25% of its total length and includes blocks as large as ~1.45 Mb with sequence identity of ~99.97% (ref. 23). In addition, many pericentromeric and subtelomeric regions are rich in dispersed segmental duplications (Fig. 5), apparently resulting from a steady bombardment of insertional translocations[46]. Although most regions of segmental duplication have now been sequenced, ~10% of them lie in the remaining gaps in the current sequence and will require further work to elucidate, as discussed above.

Protein-coding genes

A central goal of genome analysis is the comprehensive identification of all human genes. This task remains challenging, but is greatly aided by the near-complete sequence together with other improved resources (such as expanded cDNA collections, genome sequence from other organisms and better computational methods). The current version of the human gene catalogue (Ensembl 34d) contains 22,287 gene loci (with a total of 34,214 transcripts, corresponding to 1.54 transcripts per locus), consisting of 19,438 known genes and 2,188 predicted genes. These gene loci have a total of 231,667 exons, with ~10.4 exons per locus and ~9.1 exons per transcript. The total length covered by the coding exons is ~34 Mb or ~1.2% of the euchromatic genome; the untranslated regions of the transcripts are estimated to cover another ~21 Mb or ~0.7% of the euchromatic genome.

Comparison of the initial and current gene catalogues highlights the substantial improvement. Many of the earlier gene models were erroneous due to defects in the draft sequence. Examples resulting from a duplication, inversion and premature stop codon are shown in Fig. 6. The improvement can be quantified by mapping the current gene models onto the draft sequence, to determine whether they could have been accurately identified. Of the transcripts in the current gene catalogue, 58% have at least one error when mapped onto the draft sequence. For 39% of transcripts, there is at least one

Table 5 **Categories of recently arising human pseudogenes**

	Inactivating mutations in the reference sequence		
	1 mutation	>1 mutation	Total
Chimp pseudogene	8	19	27
Human pseudogene	5	0	5
Human allelic null	1	0	1
Sequence error	1	0	1
Total	15	19	34

Potential pseudogenes were identified by searching orthologous segments in human, mouse and rat (see text). Frameshifts and nonsense mutations in putative human pseudogenes were experimentally investigated by resequencing the original human BAC clones (to look for sequence errors), as well as sequencing 24 unrelated individuals (to assess whether apparent mutation is a polymorphism) and chimpanzee (to assess timing of loss of function).

exon that is absent or incorrectly ordered due to defects in the draft. For the remaining 19% of transcripts, the exons are all present and correctly ordered, but there are one or more nucleotide errors.

Automated gene annotation has now been complemented by manual annotation of most chromosomes, based on a careful review of gene structure and examination of expressed-sequence-tag (EST) and transcript evidence. Such analysis has been completed for 18 chromosomes (2, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 17, 18, 19, 20, 21, 22, X and Y; refs 17, 20–30 and http://vega.sanger.ac.uk/Homo_sapiens/), with the remainder in the press or in preparation) comprising 1.7 Gb of the euchromatic genome. Although this annotation has further improved the quality of the gene models[47] (by dealing with special cases and unusual features not yet handled in the automated programs, and resolving instances of conflicting experimental data), it has not significantly affected the total gene count for these chromosomes.

On the basis of available evidence, our best estimate is that the total number of protein-coding genes is in the range 20,000–25,000. The lower bound seems secure, based on the number of currently known genes (19,599). The upper bound is based on estimates of the number of additional genes. Despite intense automated and manual analysis using cDNA, EST and cross-species homology, only 2,188 gene predictions have been added to the known set. This predicted set is likely to represent substantially fewer than 2,000 true genes, owing to fragmentation and false predictions arising from pseudogenes. For example, the predictions tend to have fewer exons per transcript than known genes (~4.7 versus ~9.7) and to encode shorter open reading frames (~847 versus ~1,487 amino acids). On the other hand, the set is likely to be incomplete because some protein-coding genes have surely continued to escape detection. The most problematic cases would be genes that have very short open reading frames (<100 amino acids), consist of single exons or evolve very rapidly. Even if we assume that such genes comprise 10% of the total (which seems a generous overestimate, given our current understanding of the human and other genomes[48–50]), the total gene count would remain below 25,000. The range of 20,000–25,000 is also consistent with recent estimates (J. Weissenbach, unpublished) of the number of protein-coding genes based on cross-species homology (using the Exofish method[51]).

In our initial analysis of the draft sequence[15], we estimated the count of human protein-coding genes at roughly 30,000. The estimate was derived as follows. We used computational analysis to generate an initial gene catalogue with ~32,000 entries, consisting of ~15,000 known genes and ~17,000 gene predictions. We estimated that the catalogue actually corresponded to ~24,500 actual genes, based on estimates of the rate of various types of errors such as fragmentation and false positive predictions (due largely to limitations of the draft sequence, such as imperfect recognition of pseudogenes and unknown order and orientation). We then adjusted the estimate to account for the proportion of genes estimated to be absent from the initial catalogue due to incomplete coverage of the genome and imperfect computational methods, resulting in a figure of ~31,000 genes.

With the current high-quality sequence, it is now possible to revisit this earlier analysis. We directly compared the previous gene models with the current gene models, to determine whether our previous estimates of the various error rates were correct. It is clear that the main reason for the earlier overestimate is that the fragmentation rate was substantially underestimated. The fragmentation rate is defined as the average number of the previous gene models that map to the same true gene; we assessed it by mapping to the current gene catalogue. The fragmentation rates for 'known' and 'predicted' genes were estimated in our earlier paper[15] at ~1.0 and ~1.4, whereas our current analysis indicates that they should have been ~1.3 and ~1.7. This correction alone would bring our previous estimate to ~24,000. Small differences in the estimated rate of false positive and negative predictions account for the remainder of the discrepancy.

It should be emphasized that the count above refers to the count of protein-coding genes. It does not include known non-coding RNAs, such as transfer RNAs, ribosomal RNAs, small nucleolar RNAs (snoRNAs) and microRNAs[52–54]. In addition, there is evidence that the human genome gives rise to many additional RNA transcripts[55]. It is unclear whether most such transcripts have specific biological functions or reflect reproducible transcriptional noise; few contain substantial open-reading frames and thus they are unlikely to encode proteins. There is a need for reliable experimental and computational methods for comprehensive identification of non-coding RNAs.

Finally, the near-complete sequence makes it possible to undertake systematic searches for pseudogenes. Automated annotation of chromosomes has focused primarily on identifying large pseudogenes of more recent origin. Recent published studies have used more sensitive methods to detect smaller and older pseudogenes and have already identified ~20,000 processed and unprocessed pseudogenes[56]. This is surely still an underestimate, because such analysis will miss pseudogenes that are extremely old or that contain primarily untranslated regions. The total number of pseudogenes is thus likely to exceed the total number of functional genes. A particular type of pseudogene (recently arising non-processed pseudogenes) is discussed in more detail below.

Gene birth in the human lineage

The birth of new genes is of interest because it provides raw material for adaptive evolution, with extra copies of genes able to undergo functional divergence in response to positive selection. The quality and completeness of the current sequence make it possible to study this question; such analysis would have been unreliable with the earlier draft sequence, because the extensive artefactual local duplication would have given rise to many false positives.

We searched for clusters of nearby homologous genes, indicative of local gene duplication. The divergence between such genes was assessed at sites likely to be selectively neutral, by measuring the estimated substitution rate per synonymous site (K_S). We looked for nearby human gene pairs differing from one another by $K_S < 0.30$, implying that each differs from the common ancestral source gene by an average $K_S < 0.15$. This threshold corresponds roughly to duplications arising after divergence from the rodent lineage, either by recent gene duplication or perhaps recent gene conversion of older duplications (see Methods in Supplementary Information). A total of 1,183 genes exhibit such divergence from a neighbouring gene (see Methods in Supplementary Information). These genes often fall within larger clusters of paralogous genes including genes with greater divergence and reflecting older duplications. These clusters contain ~3,300 genes, and those having at least five genes involved in recent duplication events are shown in Table 4. Analysis of phylogenetic trees containing the related human and mouse genes confirms that the genes are more closely related within each species than between the two species in nearly all cases (97%), as would be expected for genes arising by duplication after the divergence of the human and rodent lineages.

The recent duplications are enriched in genes with immune and olfactory function, as well as those likely to be involved in reproductive functions. For example, the gene families encoding the pregnancy-specific beta-1-glycoprotein and choriogonadotropin beta proteins may be involved in the extended gestational period in the human lineage; the latter family is known to have expanded recently within the catarrhine primate lineage[57]. Another example is the family of cancer/testis (CT) antigen genes, which are normally expressed in the testis and are highly expressed in carcinomas[58].

The distribution of K_S values (Fig. 7) for recent duplications shows a striking excess of genes with strong similarity ($K_S \leq 0.015$), corresponding to recent events occurring ~3–4 million years ago.

There are several possible explanations for this peak. First, it may reflect a true explosion in the rate of gene duplication in the primate lineage. (The primate lineage does show an increase in the rate of dispersed segmental duplication, although it is less extreme; the rate of local duplication will need to be carefully evaluated in comparative studies.) Second, it may partly reflect on the ongoing process of gene conversion of older gene duplication events. However, we offer a third explanation: the peak primarily reflects the transient of duplicated genes that are too young relative to the characteristic time of deletion. If so, most of these new genes are destined to be culled due to lack of functional benefit. In contrast to the first explanation, this would predict that a similar peak would be seen in most mammals.

Gene death in the human lineage

Gene death is another phenomenon that sheds light on lineage-specific evolution, but which was difficult to analyse with the earlier draft sequence. To study gene death, we scanned the genome sequence for recently arising non-processed pseudogenes—that is, nearly intact human genes that appear to have recently acquired an inactivating mutation. Specifically, we examined genomic intervals bounded at each end by two consecutive genes, with each belonging to a 1:1:1 orthology triplet in the human, mouse and rat genomes and the interval containing at most 50 genes (see Methods in Supplementary Information). We then examined the Ensembl gene predictions in the corresponding intervals of the three genomes and identified instances in which the mouse and rat genomes contained 1:1 orthologues, but the human genome appeared to contain no predicted orthologous gene. In each instance, the rodent genes were aligned to the corresponding human genomic interval to look for clear evidence of a human pseudogene—that is, a highly similar sequence containing one or more inactivating mutations in its genomic sequence (see Methods in Supplementary Information). We also required that the inactivating mutation was present in any human mRNA sequences corresponding to the locus. (This analysis excludes many older pseudogenes that do not show sufficient similarity to the rodent homologues because they have substantially degenerated.)

A total of 37 candidate pseudogenes were identified, with an average of 0.8 premature stop codons and 1.6 frameshifts (Supplementary Table 1). (Similar analyses performed on the draft sequence yielded a much larger list, including many apparent inactivating mutations that were errors and were corrected in the current sequence.) We carefully examined these candidates to confirm that they did not reflect errors in the current genome sequence (by resequencing or examination of an independently finished clone) and to determine their evolutionary origin (by re-sequencing in a panel of 24 diverse humans and comparison with a draft sequence of the chimpanzee genome). Complete experimental data could be obtained for 34 cases. The identification of a pseudogene was confirmed in 33 of the 34 cases; one case was due to an error in the current sequence (Table 5). The 19 pseudogenes with two or more inactivating mutations were all found to be pseudogenes in chimpanzee as well. The 14 pseudogenes with exactly one inactivating mutation fell into the following three classes: eight pseudogenes shared with chimpanzee; five pseudogenes fixed in the human population but functional genes in the chimpanzee; and one pseudogene that is a segregating polymorphism in the human population. (In 20 cases, the inactivating mutation occurs in the final or only exon. Although this could in principle be compatible with a functional gene, the truncation removes a functionally important domain in all but one case.)

Of the 32 pseudogenes fixed in the human population, 10 are derived from olfactory receptors. Olfactory receptors thus occur prominently in both birth and death analyses, indicating a dynamic expansion and contraction of this large gene family; the net effect has been an overall significant decrease in the number of functional olfactory receptors in humans compared with rodents[59,60]. The remaining 22 recent pseudogenes include a wide variety, such as genes homologous to a cationic amino-acid transporter, a serine-threonine kinase, a calreticulin, a putative G-protein coupled receptor and a cystatin.

Discussion

The Human Genome Project marked a new approach in biomedical research, one in which the scientific community came together to characterize systematically a large domain of important biological knowledge. Because the precise scientific plan and the feasible degree of accuracy and completeness were unclear at the outset, the sequencing of the human genome proceeded in phases: a preliminary phase that developed and refined key approaches; a draft phase that yielded ~90% of the information (albeit in imperfect form); and a finishing phase reported here that yielded ~99% in high-quality form. Notably, the finishing phase required roughly equal resources of time and expense as the draft phase.

The euchromatic portion of the human genome is still not complete, with ~1% still to be determined. The issue is no longer scale, but rather the need for new approaches to understand and resolve these recalcitrant segments. Continuing efforts should be devoted towards the eventual goal of complete closure. Nonetheless, the euchromatic human genome can now be regarded as effectively known. The accuracy and completeness of the current near-complete human genome sequence has important consequences for biomedical research. It allows systematic searches for the causes of disease—for example, to find all key heritable factors predisposing to diabetes or somatic mutations underlying breast cancer—with confidence that little can escape detection. It facilitates experimental tools to recognize cellular components—for example, detectors for mRNAs based on specific oligonucleotide probes or mass-spectrometric identification of proteins based on specific peptide sequences—with confidence that these features provide a unique signature. It allows sophisticated computational analyses—for example, to study genome structure and evolution—with confidence that subtle results will not be swamped or swayed by noisy data. At a practical level, it eliminates tedious confirmatory work by researchers, who can now rely on highly accurate information. At a conceptual level, the near-complete picture makes it reasonable for the first time to contemplate systems approaches to cellular circuitry, without fear that major components are missing.

The HGP provides an essential foundation for the sequencing and analysis of additional large genomes. With the experience gained from the human genome, it has already become scientifically and economically feasible to produce draft genome sequence from many vertebrates, which will be a crucial tool for identifying the functional elements in the human genome through comparative analysis. Ultimately, we believe that such projects should aim higher to produce genome sequence with even greater accuracy and completeness. This will require digesting the diverse experience from the finishing phase of human sequencing and selecting a subset of techniques that can be most efficiently streamlined and scaled up to improve accuracy and completeness of genome sequence. A good example is the systematic closure of gaps by primer-directed walking on fosmid templates covering each gap, which may be able to close the vast majority of gaps in a draft sequence in an automated fashion.

More generally, the HGP demonstrates the tremendous potential value of coordinated projects to create community resources to propel biomedical research. Key challenges that lie ahead[61] include: (1) systematic identification of all genetic polymorphisms carried in the human population, to facilitate the study of their association with disease; this will require comprehensive study of hundreds to thousands of human genomes. (2) Systematic identification of all functional elements in the human genome, including genes, proteins, regulatory controls and structure elements; this will require

comparative analysis with many additional mammalian genomes and systematic application of diverse experimental techniques. (3) Systematic identification of all the 'modules' in which genes and proteins function together; this will require comprehensive study and improved interpretation of expression, localization and interaction in a temporal and spatial context. Absolute completeness will be elusive but, as with the HGP, obtaining the substantial majority of the information will greatly accelerate the pace of biomedical research in thousands of laboratories. □

Received 29 July; accepted 7 September 2004; doi:10.1038/nature03001.

1. NIH/ CEPH Collaborative Mapping Group. A comprehensive genetic linkage map of the human genome. *Science* **258,** 67–86 (1992).
2. Gyapay, G. *et al.* The 1993–94 Genethon human genetic linkage map. *Nature Genet.* **7,** 246–339 (1994).
3. Murray, J. C. *et al.* A comprehensive human linkage map with centimorgan density. *Science* **265,** 2049–2054 (1994).
4. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
5. Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270,** 1945–1954 (1995).
6. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
7. International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
8. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380,** 149–152 (1996).
9. Gregory, S. G. *et al.* A physical map of the mouse genome. *Nature* **418,** 743–750 (2002).
10. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
11. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1457 (1997).
12. Goffeau, A. *et al.* Life with 6,000 genes. *Science* **274,** 546–567 (1996).
13. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282,** 2012–2018 (1998).
14. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
15. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
16. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
17. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **401,** 489–495 (1999).
18. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
19. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3,** 0079.1–0079.14 (2002).
20. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
21. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
22. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
23. Skaletsky, H. *et al.* The male-specific regions of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
24. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
25. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
26. Dunham, A. *et al.* The DNA sequence and analysis of human chromosome 13. *Nature* **428,** 522–528 (2004).
27. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
28. Humphray, S. J. *et al.* DNA sequence and analysis of human chromosome 9. *Nature* **429,** 369–374 (2004).
29. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429,** 375–381 (2004).
30. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431,** 268–274 (2004).
31. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).
32. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
33. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.* **31,** 34–37 (2003).
34. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99,** 16899–16903 (2002).
35. Furey, T. S. *et al.* Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing. *Genome Res.*(in the press).
36. Riethman, H. C., Moyzis, R. K., Meyne, J., Burke, D. T. & Olson, M. V. Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast artificial chromosome vector. *Proc. Natl Acad. Sci. USA* **86,** 6240–6244 (1989).
37. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Rev. Genet.* **5,** 345–354 (2004).
38. Lai, Z. *et al.* A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.* **23,** 309–313 (1999).
39. She, X. *et al.* The structure and evolution of centromeric transition regions within the human genome. *Nature* **430,** 857–864 (2004).
40. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* (in the press).
41. Nilsson, M. *et al.* Padlock probes reveal single-nucleotide differences, parent of origin and *in situ* distribution of centromeric sequences in human chromosomes 13 and 21. *Nature Genet.* **16,** 252–255 (1997).
42. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18,** 74–82 (2002).
43. Johnson, M. E. *et al.* Positive selection of a novel gene family during the emergence of humans and great apes. *Nature* **413,** 514–519 (2001).
44. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14,** 789–801 (2004).
45. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway Rat. *Genome Res.* **14,** 493–506 (2004).
46. Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10,** 2215–2223 (2001).
47. Collins, J. E. *et al.* Re-evaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13,** 27–36 (2003).
48. Cliften, P. F. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11,** 1175–1186 (2001).
49. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301,** 71–76 (2003).
50. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423,** 241–254 (2003).
51. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
52. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
53. Storz, G. An expanding universe of noncoding RNAs. *Science* **296,** 1260–1263 (2002).
54. Szymanski, M., Erdmann, V. A. & Barciszewski, J. Noncoding regulatory RNAs database. *Nucleic Acids Res.* **31,** 429–431 (2003).
55. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).
56. Torrents, D., Suyama, M., Zdobnov, E. & Bork, P. A genome-wide survey of human peudogenes. *Genome Res.* **13,** 2559–2567 (2003).
57. Maston, G. A. & Ruvolo, M. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. *Mol. Biol. Evol.* **19,** 320–355 (2002).
58. Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J. & Chen, Y.-T. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol. Rev.* **188,** 22–32 (2002).
59. Glusman, G., Yanai, I., Rubin, I. & Lancet, D. The complete human olfactory subgenome. *Genome Res.* **11,** 685–702 (2001).
60. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
61. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422,** 835–847 (2003).
62. Lee, C., Weverick, R., Fisher, B. B., Furguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100,** 291–304 (1997).
63. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88,** 7474–7476 (1991).
64. Madan, K. & Bobrow, M. Structural variation in chromosome no. 9. *Ann. Genet.* **17,** 81–86 (1974).
65. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
66. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70,** 83–100 (2002).
67. Loftus, B. J. *et al.* Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. *Genomics* **60,** 295–308 (1999).
68. Gordon, D., Desmarais, C. & Green, P. Automated finishing with autofinish. *Genome Res.* **11,** 614–625 (2001).
69. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101,** 1916–1921 (2004).
70. McMurray, A. A., Sulston, J. E. & Quail, M. A. Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res.* **8,** 562–566 (1998).
71. Heiner, C. R., Hunkapiller, K. L., Chen, S. M., Glass, J. I. & Chen, E. Y. Sequencing multimegabase-template DNA with BigDye terminator chemistry. *Genome Res.* **8,** 557–561 (1998).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank D. Leja for graphic design and production of the figures. We would also like to thank the many dedicated support staff at the sequencing centres and funding agencies. In addition to finished sequence produced by the IHGSC between the time of publication of the draft human genome and April 2003, Build 35 contains some significant published finished sequence from other centres as listed in Table 1—for this we would like to acknowledge M. Adams, B. Roe and G. Evans. Build 35 also contains a small number of individual finished deposited accessions from a variety of other groups, which may or may not have been published; we would like to acknowledge all of this work. We acknowledge G. Sisk for help in preparing the manuscript. This work was supported by The Wellcome Trust; The US National Institutes of Health; The US Department of Energy; The Ministry of Education, Culture, Sports, Science and Technology, Japan; The Federal German Ministry of Education, Research, and Technology; Projektträger Biologie, Energie, Umwelt des BMBF und BMWT; the Max-Planck-Society; Deutsche Forschungsgemeinschaft; Thüringer Ministerium für Wissenschaft, Forschung, und Kunst; The Medical Research Council (UK); European Commission, Directorate Science, Research and Development; Chinese Academy of Sciences, Ministry of Science and Technology, National Natural Science Foundation of China.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.S. Collins (fc23a@nih.gov), E. S. Lander (lander@broad.mit.edu), J. Rogers (jrh@sanger.ac.uk) or R. H. Waterston (waterston@gs.washington.edu). The sequence described here has been deposited in public databases, with the 24 human chromosomes having accession numbers NC000001 to NC000024.